Software Technology to Enable Reliable High-Performance Distributed Disk Arrays

Michael S. Warren, Chris L. Fryer, M. Patrick Goda, and Ryan Joseph (T-6); ***mswarren@lanl.gov***

It is currently possible to construct a single-node RAID storage system with a 2 terabyte capacity using commodity serial-ATA hard disk drives for less than $2K. Within two years, a cluster of such systems (a distributed disk array) will be able to provide over a petabyte of storage for less than $1M. Obtaining reliability and good performance from such a system is the focus of our project.

Over the past year, we have established a number of testbed systems containing over 50 terabytes of storage. Roughly half of this amount is in the 288-processor Space Simulator Beowulf cluster [1, 2], with the remainder in a variety of RAID systems. Significant progress to date includes detailed failure statistics on a variety of disk drives, performance benchmarks on a number of different systems, and modifications to the Linux Network Block Device (NBD) driver to support RAID-5 arrays across multiple cluster nodes.

As an introduction, the advent of commodity microprocessors with adequate floating-point performance and low priced fast Ethernet switches contributed to the emergence of Beowulf clusters in the mid-90s. We are currently poised for a similar advance in distributed disk arrays (DDAs), due to the dramatic decline in the price of commodity disk drives.

The cost per gigabyte for 7200 RPM IDE disk drives is currently less than $1. Several groups have demonstrated fault-tolerant terabyte RAID servers for a total cost of under $2K per terabyte. Used in a parallel cluster environment, multiterabyte disk arrays with achievable read/write bandwidths that greatly exceed available gigabit local- and wide-area networking technology are possible. Additionally, the greater CPU/storage ratio in a DDA offers techniques that are not possible in traditional RAID arrays.

The NBD exists as a standard part of the Linux kernel and provides the functionality required for network attached storage. An NBD is "a long pair of wires." It makes a remote disk on a different machine act as though it were a local disk on your machine. It looks like a block device on the local machine where it is typically going to appear as /dev/nda. The remote resource does not need to be a whole disk or even a partition. The NBD system works by emulating a block device on the client side, while actual requests to that device are passed over the network to the true block device or file on one or many NBD servers. The NBD consists of two user-space programs and one kernel-space module: the two user-space programs are a client and server, respectively, while the module loaded into any kernel that wishes to be a client for NBD.

Figure 1— AICIPC Hot-Swap Chassis (Parallel ATA and Serial ATA).

Figure 2—
Serial ATA solutions.

Qty.	Price	Ext.	Description
1	220	220	Intel P4 Processor 3.2GHz, 800MHz FSB
8	385	3080	IBM Hitachi Deskstar 400GB SATA 7200RPM
1	530	530	3ware 9500-8 RAID card
1	322	322	TYAN GC-SL S2727G2N with Dual GigE and PCI-X
2	79	158	Corsair 2x512MB DDR400 PC3200 memory
1	150	150	Supermicro 5 Bay Hot-Swapable SATA HDD Enclosure
1	75	75	Mid-Tower case, 4x5.25 exposed, 4x3.5 int., extra fans
1	94	94	Enermax EG465P-VE (FCA) 431W Power Supply
1	80	80	Assembly
Total		**$4709**	**$1.47 per Gbyte**

Table 1—
Mid-tower storage system pricing, November 2004.

We evaluated a number of systems to quantify the performance of single disks, as well as different types of hardware and software RAID-5 systems. Our initial results indicate the best price/performance system uses Serial ATA disks with a 3ware controller and Linux software RAID.

We investigated a number of storage systems to determine the optimal "building block" for larger systems. Our first reference design is listed in Table 1, which has a storage capacity of 2 terabytes at a cost of about $4K. We purchased 10 such systems, which will be the testbed hardware for further software development.

[1] M.S. Warren, C.L. Fryer, and M.P. Goda, "The Space Simulator," in *Proceedings of CWCE '03*, San Jose, 2003.
[2] M.S. Warren, C.L. Fryer, and M.P. Goda, "The Space Simulator: Modeling the universe from Supernovae to Cosmology," in *Proceedings of the ACM/IEEE SC2003 Conference*, (ACM Press, New York), 2003.